A new diagnostic to assess information for variance parameter estimation in Multi-Environment Trials

Chris Lisle[1,2]
clisle@uow.edu.au

Alison Smith[1,2], Carole Birrell[2], Brian Cullis[1,2]

[1]Centre for Biometrics and Data Science for Sustainable Primary Industries (CBADS-SPI)
[2]National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong (UOW)

Australasian Plant Breeding Conference, Gold Coast, 09 - 11 May 2022

May 11, 2022

CBADS-SPI
CENTRE FOR BIOMETRICS AND DATA SCIENCE FOR SUSTAINABLE PRIMARY INDUSTRIES

UNIVERSITY OF WOLLONGONG AUSTRALIA

Overview of talk

1. Key motivation
2. New diagnostic
3. Motivating Durum dataset
4. Simulation Studies to Investigate the Performance of the Diagnostic
5. Conclusion

Acknowledgements

- Research presented here is part of my PhD research.

  Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2021). Information Based Diagnostic for Genetic Variance Parameter Estimation in Multi-Environment Trials. Frontiers in Plant Science, 12, 2856.

- PhD title: Information based diagnostics for the optimal construction of Multi-environment trial datasets.
- Special thanks to my supervisors Alison Smith, Carole Birrell, and Brian Cullis.

MET dataset analysis

- Selection of superior varieties is a result of the data analysis from a series of plant variety trials at a number of locations and possibly over several years, which are known as Multi-Environment Trials (METs).
- A 1-stage Factor Analytic Linear Mixed Model (FALMM) (Smith et al., 2001) to model Variety by Environment (VE) effects is considered to be the gold standard.
- Substantial literature demonstrates increased genetic gains from the use of FALMM.
- Ancestral information of the varieties grown in the dataset is often available through a 'pedigree file', allowing the partition of the VE effects into additive and non-additive.
- A FALMM that incorporates pedigree information has also been shown to improve selection accuracy (for example Oakey et al., 2007).

Factor Analytic Linear Mixed Model

- A 1-stage analysis so uses individual plot data combined across all trials in the MET
- The LMM comprises fixed and random effects and errors
- The random effects include
  - genetic: VE effects partitioned into additive (associated with pedigree) and non-additive
  - non-genetic: effects associated with trial designs (eg. replicate blocks) and extraneous variation (rows/columns)
- The errors relate to plot to plot variation within trials

Factor Analytic Linear Mixed Model

- The random effects and errors require variance models
  - genetic: FA model for between environment additive genetic variance matrix; FA model for between environment non-additive genetic variance matrix
  - non-genetic: often simple variance components
  - errors: spatial models for each trial (spatial variances, row/column auto-correlations)

- Main focus of MET analysis is accurate prediction of genetic (VE) effects
- Predictions of random effects, including genetic, depend on the variance parameters (genetic, non-genetic and error)
- If we knew the variance parameters we would obtain best linear unbiased predictions (BLUPs) of the genetic effects
- But the variance parameters are unknown so we use residual maximum likelihood (REML, Patterson & Thompson, 1971) estimates obtained from fitting the FALMM

- Use of variance parameter estimates leads to *empirical* best linear unbiased predictions (EBLUPs) of the genetic effects
- EBLUPs are not as accurate as BLUPs
- The (loss in) accuracy for EBLUPs depends on the accuracy/reliability of the REML estimates of the variance parameters (Sales & Hill, 1976)

Key motivation for my research

- Due to the evaluation process, MET datasets are generally (highly) unbalanced.
  $\rightarrow$ most varieties are not present in all environments.
- "Variety Connectivity": the number of varieties in common between environments.
- We observe varying levels of variety connectivity when forming appropriate MET datasets such as the contemporary group methodology of Smith et al. (2021).
  - Dataset is often formed across years and breeding stages.
  - The $\mathcal{A}$-optimality criterion is used to quantify different datasets since this aligns with minimising the probability of an incorrect selection decision (Bueno Filho and Gilmour, 2003).
  - However, this criterion assumes known variance parameters, whereas in practice they are required to be estimated.
- We thought variety connectivity was a key driver of the reliability of genetic variance parameter estimates and that this in turn affected the reliability of predictions of VE effects. **Attributed to the amount of information, or lack there of**.

Variety connectivity

- Variety Connectivity
  - Strong connectivity – many/all varieties in common.
  - Poor connectivity – few (none!) varieties in common.
- Easy to visualise via tables and figures.
- We sometimes exclude environments with poor connectivity from our MET datasets.
- We often suggest ad hoc rules of thumb in regards to the number of varieties that should be present between environments and/or over years.

Investigation of variety connectivity

- Simulation study to investigate the effect of variety connectivity on independent variety effects (Lisle et al., 2018; Lisle et al., 2019).
  - Varied the number of common varieties (m) between trials (1:m).
- Results showed a response in reliability of the Empirical Best Linear Unbiased Predictions (E-BLUPs) over levels of variety connectivity.
  - Correlation between the true (known) and predicted values (those estimated).
- More happening than just the number of varieties in common affecting the performance of variety reliabilities.
  - Difference in the responses for connected and not connected varieties.
  - Increases in reliability values for larger trial sizes.
  - There were some interesting/unexpected results!

Links with fundamental concepts in experimental design theory

- We want to understand about the reliability of genetic variance parameter estimates.
- We require a diagnostic that outperforms the existing variety connectivity approach and can also account for genetic relationships.
- In model-based design we seek designs that minimises/maximises a chosen criteria.
- These include the so-called alphabet series of optimality criteria $(\mathcal{A}, \mathcal{D}, \mathcal{E}, \dots)$.
- $\mathcal{D}$-value optimality minimises the variance of treatment estimates. This is calculated as the determinant of the variance matrix of treatment effects.
- The $\mathcal{D}$-value measure is therefore applicable to our problem where we want to assess the **variance of estimated variance parameters**.

$\mathcal{D}$-value diagnostic of Lisle et al. (2021)

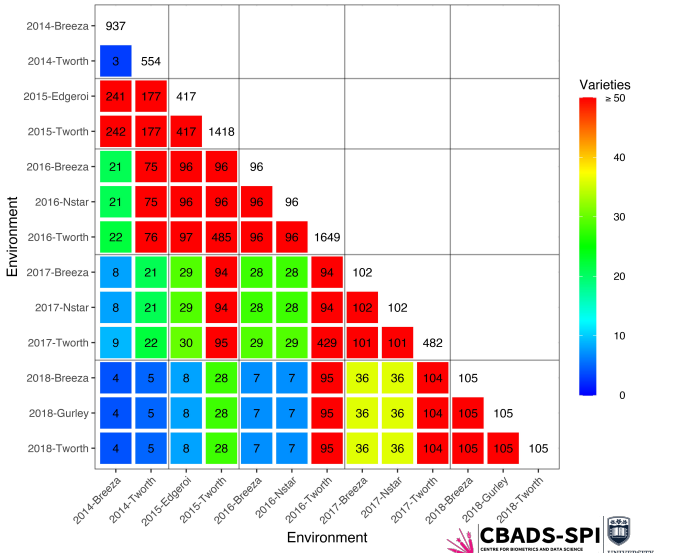
- Calculated as the log-determinant of the partition elements of the inverse expected information matrix for the genetic variance parameters of interest.
  - See Equations (11) and (12) in Lisle et al. (2021).
  - **Warning:** Some heavy algebra!
- We can look at the overall $\mathcal{D}$-value or individual environments ($\mathcal{D}_j$-values).
- Lower diagnostic $\mathcal{D}$-values represent higher levels of information to estimate the parameters of interest. **Lower values are better**.
- Can be investigated over a wide range of models. e.g those with or without genetic relationships information.

Motivating Durum MET dataset

- Selection decisions for Stage 3 (S3) varieties for 2018.
- CG: environments sown between 2014 and 2018 from breeding stages Stage 1 (S1) to S3
- 13 environments, 40 trials, 3,708 varieties, 9,786 plots, 6,168 VE combinations.
  - An environment can comprise of multiple trials which may include different selection stages.
  - We use environment to model VE interaction.
- Number of varieties per environment ranged from 96 to 1,649, with a median of 105.
- Pedigree information was given with 3,959 records.
  - Information for the 3,708 varieties that were grown in the environments and 251 parental varieties that were not grown in the environments.

Variety Connectivity

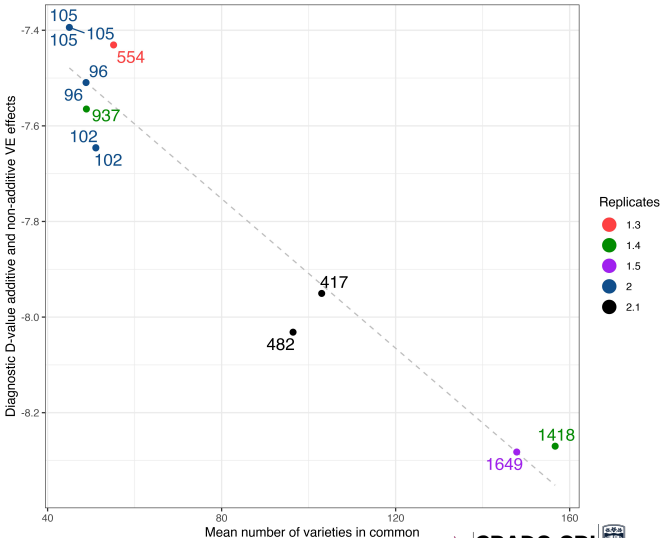
- Can be viewed via figures/tables.
- With ancestral information through the pedigree file, we can also inspect the parental connectivity.
- But there are also common grandparents, cousins, great grandparents etc.
- However, we do not have methods which fully quantify all relationships.



Putting $\mathcal{D}_j$-value into practice: with pedigree information

- We use representative parameter values from historical analyses.
  - 80% additive variance.
  - Between environments correlations of 0.8 for Add/N_Add VE effects.

Environment	$\mathcal{D}_j$-value
2016-Tworth	-8.28
2015-Tworth	-8.27
2017-Tworth	-8.03
2015-Edgeroi	-7.95
2017-Breeza	-7.65
2017-Nstar	-7.65
2014-Breeza	-7.56
2016-Breeza	-7.51
2016-Nstar	-7.51
2014-Tworth	-7.43
2018-Breeza	-7.39
2018-Gurley	-7.39
2018-Tworth	-7.39
n_{κ_g}	182



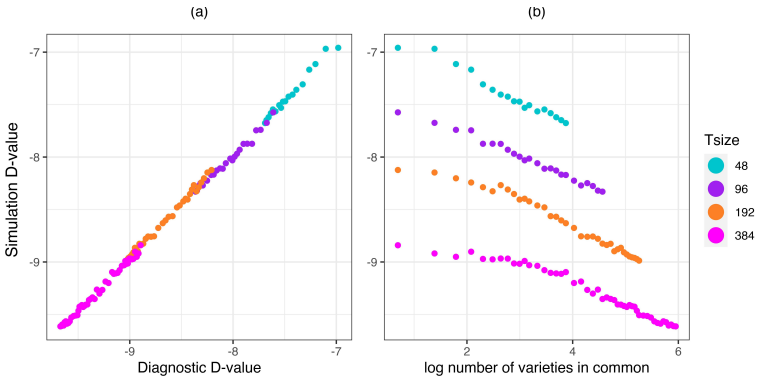
CBADS-SPI
CENTRE FOR BIOMETRICS AND DATA SCIENCE FOR SUSTAINABLE PRIMARY INDUSTRIES

UNIVERSITY OF WOLLONGONG AUSTRALIA

Simulation Studies to Investigate the Performance of the Diagnostic

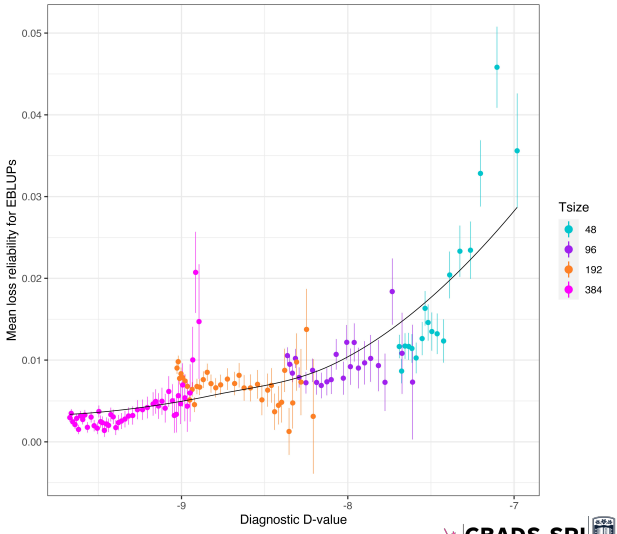
Four trial size (tsize) with $p = 2$ environments, varying the number of varieties in common.

- Genetic relatedness from the durum example.
- $N = 2000$.
- (a) Simulated vs diagnostic $\mathcal{D}$-values show agreement.
- (b) Decreasing linear relationship with (log) variety connectivity, but only within tsize.



Associated loss in EBLUP reliabilities

- Loss is taken as the difference in reliability due to the estimation of variance parameters.
- Smaller trial sizes have a larger associated loss in EBLUP reliabilities.
- The loss is well predicted by the diagnostic $\mathcal{D}$-value.
- This is seen both within and across trial sizes.



Conclusion

- The $\mathcal{D}$-value diagnostic is to be applied to a MET dataset prior to analysis.
- Environments with high $\mathcal{D}$-values may contribute insufficient information for genetic variance parameter estimates, so their inclusion in the MET dataset should be considered.
- Our diagnostic encapsulates not just variety connectivity but also other structural features such as trial size, replication and genetic relatedness.
- Even in the case of 100% connectivity, the smallest trial sizes resulted in large $\mathcal{D}$-values which then translated to substantial losses in the reliability of VE predictions.

Key references

- **Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2021). Information Based Diagnostic for Genetic Variance Parameter Estimation in Multi-Environment Trials. Frontiers in Plant Science, 12, 2856.**
- Smith, A., Ganesalingam, A., Lisle, C., Kadkol, G., Hobson, K., & Cullis, B. R. (2021). Use of Contemporary Groups in the Construction of Multi-Environment Trial Datasets for Selection in Plant Breeding Programs. Frontiers in Plant Science, 11, 2325.
- Lisle, C. (2019). Varietal Connectivity: does it affect the accuracy of variety predictions from factor analyltic multi-environment trial analyses? In Wheat Breeding Assembly, August 20-22 2019, Adelaide Oval.
- Lisle, C., Smith, A., Birrell, C., & Cullis, B. R. (2018). Varietal connectivity: Does it affect the accuracy of results from a multi-environment trial analysis? In Australasian Applied Statistics Conference. Rotorua, New Zealand, 3-7 December 2018.

References

- Smith, A., Cullis, B. R., & Thompson, R. (2001). Analyzing Variety by Environment Data Using Multiplicative Mixed Models and Adjustments for Spatial Field Trend. Biometrics, 57(4), 1138–1147.
- Oakey, H., Verbyla, A., Cullis, B. R., Wei, X., & Pitchford, W. S. (2007). Joint modelling of additive and non-additive (genetic line) effects in Mulit-environment trials. Theoretical and Applied Genetics, 114, 1319–1332.
- Patterson, H. D., & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. Biometrika, 58, 545–554.
- Sales, J., & Hill, W. G. (1976). Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. Animal Production Science, 23(01), 1–14.
- Bueno Filho, J., & Gilmour, S. (2003). Planning incomplete block experiments when treatments are genetically related. Biometrics, 59(2), 375–381.